Ambiguity Resolution while Mapping Free Text to the UMLS Metathesaurus

Thomas C. Rindflesch and Alan R. Aronson
National Library of Medicine
Bethesda, MD 20894

We propose a method for resolving ambiguities encountered when mapping free text to the UMLS® Metathesaurus.® Much of the research in medical informatics involves the manipulation of free text. The Metathesaurus contains extensive information which supports solutions to problems encountered while processing such text. After discussing the process of mapping free text to the Metathesaurus and describing the ambiguities which are often the result of such mapping, we provide examples of rules designed to eliminate mapping ambiguities. These rules refer to the context in which the ambiguity occurs and crucially depend on semantic types obtained from the Metathesaurus. We have conducted a preliminary test of the methodology and the results obtained indicate that the rules successfully resolve ambiguity around 80% of the time.

INTRODUCTION

As automated methods in medical informatics mature, researchers are increasingly addressing the problems inherent in manipulating free text. Due to the complexity of natural language, such processing poses a particular challenge to system developers. The Unified Medical Language System® (UMLS) ([1]) provides extensive support for processing natural language.

Several studies ([2], [3], and [4], for example) discuss projects which exploit the UMLS Metathesaurus in natural language processing. In addition to its value for such research, [1] summarizes research using UMLS for a variety of purposes involving the manipulation of text, including information retrieval, indexing, and data creation applications.

In order to effectively use the information contained in the 4th (1993) Experimental Edition of the Metathesaurus it is first necessary to map the text being processed to Metathesaurus concepts. Many researchers have proposed various methods for such mapping. (The early projects map to the MeSH® vocabulary, but the principles involved are identical to those involved in mapping to the Metathesaurus.) Some examples are [5] (mapping to MeSH), [6] (mapping to MeSH), [7], [8], [9], [10], [11], and [12].

Regardless of the method used for mapping to the Metathesaurus, ambiguous mappings result. Such mappings, which occur when a text phrase maps correctly to more than one Metathesaurus concept, have to be resolved before further processing can be reliably pursued. In this paper, after discussing the process of mapping free text to Metathesaurus concepts and elucidating the resulting ambiguity, we describe a pilot study which investigates an approach to resolving such ambiguity.

MAPPING PHRASES IN FREE TEXT TO THE UMLS METATHESAURUS

An example of a general and robust algorithm for mapping to a controlled vocabulary is described in [6]. This algorithm has a number of characteristics which ensure that the concepts identified accurately represent the source text. The algorithm first identifies noun phrases in the input text and then maps to concepts within each noun phrase. It further produces morphological variants, and deals with various kinds of partial matches, including permutation of words, synonymy, intervening unimportant words, partial matches, and complex matches.

We have developed a program for mapping free text to the Metathesaurus which has most of the desirable characteristics found in [6], but which differs considerably in implementation and in our treatment of partial and complex matches. Our program, which is described in more detail in [4], first identifies simple noun phrases in free text. This syntactic analysis relies on the SPECIALIST lexicon ([13]) and the Xerox Part-of-Speech Tagger ([14]). The following examples demonstrate the crucial characteristics of our mapping program as it identifies Metathesaurus concepts for each noun phrase.

We employ intensive variant generation, which, in addition to accommodating purely string-based variants, such as upper and lower case distinctions and inflectional variants, establishes a relationship

0195-4210/94/$5.00

between variants based on derivational morphology, synonymy, and abbreviation. For example, although the term *renal transplant* does not occur in the Metathesaurus, our variant generation determines that *renal* is a synonym of *kidney* and thus the text *renal transplant* maps to the Metathesaurus term "KIDNEY TRANSPLANT", one of the terms for the concept "Kidney Transplantation". Similarly, *thermogram* is not in the Metathesaurus; however, morphological variant generation allows this term to map to "Thermography". Finally, our treatment of abbreviations and acronyms allows us to map *ICU* to "Intensive Care Unit".

We also allow complex matches, in which more than one Metathesaurus concept represents the text of a noun phrase. For example, the noun phrase in (1a) maps to the Metathesaurus concepts (1b) and (1c).

(1) a. digoxin overdose

b. "Digoxin"

c. "Overdose"

It should be noted that the mapping algorithm does not itself specify the relationship between the Metathesaurus concepts in a complex match; further processing is required to determine the relationship between concepts in complex matches.

For those cases in which only part of the noun phrase has a mapping to the Metathesaurus, we distinguish between instances in which the head is involved in mapping (2), and those in which it is not (3). (The head of a noun phrase is the rightmost noun in the structure.)

(2) a. liquid crystal thermography

b. "Thermography"

(3) a. cochlear implant subjects

b. "Cochlear Implant"

A further sub-type of partial match involves a mapping to Metathesaurus concepts which do not contiguously cover the text of the noun phrase, as in (4).

(4) a. adjuvant-induced arthritis

b. "Arthritis, Adjuvant"

(5) illustrates the problem of ambiguous terms in the Metathesaurus. The noun phrase in (5a) has a complete mapping to the Metathesaurus as indicated in (5b-d). However, *dialysis* maps to the two concepts shown, where the corresponding semantic types are given in brackets to indicate the two meanings of *dialysis*.

(5) a. lymph dialysis

b. "Lymph"

c. "Dialysis <1>" ['Natural Phenomenon or Process']

d. "Dialysis <2>" ['Therapeutic or Preventive Procedure']

AMBIGUOUS MAPPINGS

Regardless of the effectiveness of the algorithm employed for mapping free text to concepts in the Metathesaurus, ambiguities will result. Ambiguous mappings to the Metathesaurus fall into two general categories: those caused by variant generation (this may be due to morphological variants, synonyms, or abbreviations), and those caused by ambiguity inherent in a Metathesaurus concept itself, as illustrated in (5) above.

With regard to variant generation, morphology together with synonymy conspire to produce multiple mappings which must be resolved in a particular context. For example, these phenomena cause the word *fundamental* to map infelicitously to the Metathesaurus concepts "Foundation" (with semantic type 'Organization') and "base" (with semantic type "Inorganic Chemical"). Abbreviations cause particular problems due to the fact that they often have several expansions, none of which are semantically related to each other. For example, the single letter *c* matches Metathesaurus concepts including: "Carbon", "Complement", Cytidine", and "Cytosine".

Ambiguous words and phrases are distinguished in the Metathesaurus either by integers in angled brackets or by a note in parentheses. The ambiguous word *dialysis,* for example, is represented by the preferred term "Dialysis <1>", which has semantic type 'Natural Phenomenon or Process' and by the preferred term "Dialysis <2>", which has semantic type 'Therapeutic or Preventive Procedure'. *Inhibition* is represented as "Inhibition (Psychology)" with semantic type 'Mental Process' and as "Inhibition <2>" with semantic type 'Molecular Function'.

Ambiguities, of whatever type, have to be resolved if the Metathesaurus concepts obtained by the mapping algorithm are to accurately support further processing of the input text. We have decided to first address the type of ambiguity which is due to ambiguous concepts in the Metathesaurus itself. The technique being

developed can then serve as the basis for resolving ambiguity due to variant generation. (Also see [2] for an example of ambiguity resolution in the context of a natural language processing system.)

AMBIGUITY RESOLUTION

The type of ambiguity exemplified by ambiguous Metathesaurus concepts is often referred to as word sense ambiguity. The general principle which supports the resolution of this ambiguity is the notion that a particular sense of a term occurs in a definable textual context. Beginning with [15] a number of researchers in computational linguistics have proposed various systems for exploiting contextual information for the purposes of word sense disambiguation (for example [16], [17], and [18]). We would like to take advantage of the insights these systems offer with regard to the general approach to word sense disambiguation, especially with regard to what kinds of information can contribute to the disambiguation and the general ways in which this information can be exploited.

A further consideration is how much context has to be specified for effective ambiguity resolution. That is, the context might be the sentence in which the word occurs, or it might be the paragraph, or it might be some larger text unit, for example the entire document in which the ambiguous term occurs.

As a pilot study, we have implemented a word sense disambiguation algorithm in a Prolog program and have tested it on the NLM Test Collection ([19]). Our system has been influenced in particular by [18]; however, in this prototype system, we limit the context used for disambiguation to that occurring in the sentence in which the ambiguous term occurs. Within this context, the rules which resolve ambiguity may refer either to the presence of patterns of particular words or to patterns of UMLS semantic types associated with the Metathesaurus concepts which constitute the mapping of the sentence in which the ambiguous term occurs. Either of these pattern types may be defined as occurring in a particular syntactic structure. In the discussion that follows we provide examples of the process of ambiguity resolution using partial, informal statements of the rules involved.

The disambiguation process is driven by rules which are associated with semantic types. Each semantic type has associated with it a disambiguation rule which specifies the evidence that supports selection of this semantic type. Upon selection of a semantic type, ambiguity is resolved in that the Metathesaurus concept associated with that type is selected and the other candidates are rejected. An important characteristic of the entire approach is that it is probabilistic; the successful application of a rule in favor of a particular semantic type indicates that there is a certain likelihood that the ambiguity should be resolved in favor of that semantic type, but the determination is not categorical.

As an example of the application of the disambiguation rules first note that *immunology* ambiguously maps to the Metathesaurus concepts shown in (6).

(6) a. "immunology <1>" ['Biologic Function']

b. "Immunology <2>" ['Biomedical Occupation or Discipline']

c. "Immunology <3>" ['Laboratory Procedure']

One of the rules for the semantic type 'Laboratory Procedure' is:

(7) Evidence in favor of 'Laboratory Procedure':

> One of the following list of words occurs to the right of the ambiguous concept: classify, indicate, procedure, reveal, show, analysis, experiment, finding, method, technique.

In the text given in (8), rule (7) applies to select semantic type 'Laboratory Procedure' and its associated concept "Immunology <3>" since the word *analysis* follows the word in the text which is involved in the mapping ambiguity. (In this and the following examples, the textual material which is involved in a mapping ambiguity is underlined, and the textual context which contributes to the resolution of the ambiguity is given in bold type.)

(8) Immunological **analysis** of the released fibronectin indicated that LTA was the only surface component which could be detected as a soluble complex with the released fibronectin.

The following rule describes some of the evidence which supports selection of the type 'Biologic Function'. (Reference to a semantic type in a rule is to be interpreted as also referring to all the children of that semantic type in the UMLS Semantic Network.)

(9) Evidence in favor of 'Biologic Function':

> A prepositional phrase occurs to the right of the ambiguous concept, the preposition is *in* or *of*, and the head of the object of the preposition maps to a concept which has the semantic types 'Plant' or 'Animal'. The prepositional phrase occurs "close" to the

ambiguous concept but need not be immediately contiguous.

OR

A phrase which maps to a concept having the semantic type 'Disease or Syndrome' occurs to the right of the ambiguous concept.

In the following examples, rule (9) chooses the correct mapping for *immunology*. In both (10) and (11) this is the concept "immunology <1>", which has the semantic type 'Biologic Function'. In (10) the ambiguous term is followed by one which has the semantic type 'Animal'. In (11) the contextual evidence supporting selection of 'Biologic Function' is a term having the semantic type 'Disease or Syndrome'.

(10) The immunological responses **of owl monkeys** to L. b. panamensis were similar in many respects to those observed in humans with localized cutaneous leishmaniasis.

(11) This nonhuman primate model should be useful for future studies involving the immunology and chemotherapy of **cutaneous leishmaniasis**.

There is a general principle that if the evidence which could support a particular semantic type does not in fact occur in the text, then that semantic type is disfavored and the alternatives are favored. *Imipramine* (as shown in (12)) is one of the Metathesaurus terms which belong to the ambiguity class having one semantic type 'Laboratory Procedure', and one or more additional semantic types which are children of 'Substance' in the Semantic Network.

(12) a. "Imipramine <1>" ['Organic Chemical', 'Pharmacologic Substance']

b. "Imipramine <2>" ['Laboratory Procedure']

In examples (13) and (14), although the rule for 'Laboratory Procedure' has a chance to apply (since that semantic type occurs associated with one of the possible mappings), there is no evidence to support selection of 'Laboratory Procedure'. Thus, this semantic type is eliminated in favor of the alternative semantic types ('Organic Chemical' and 'Pharmacologic Substance') and "Imipramine <1>" is selected to resolve the ambiguity.

(13) Moreover, imipramine, an inhibitor of protein kinase C, had little effect.

(14) Fluoxetine has overall therapeutic efficacy comparable with imipramine, amitriptyline and doxepin in patients with unipolar depression treated for 5 to 6 weeks, although it may be less effective than tricyclic antidepressants in relieving sleep disorders in depressed patients.

Additional rules might enhance the evidence in support of either 'Organic Chemical' or 'Pharmacologic Substance' but could not contradict the elimination of 'Laboratory Procedure'. Alternatively, after the elimination of 'Laboratory Procedure', it could be the case that there is also no evidence to support either 'Organic Chemical' or 'Pharmacologic Substance'. In such a case the ambiguity must be left unresolved.

TESTING THE METHODOLOGY

In order to determine the viability of our general approach to ambiguity resolution, we conducted a pilot study limited to instances of *immunology* in the NLM Test Collection. We extracted 110 sentences containing instances of morphological variants of *immunology*. As noted above *immunology* maps ambiguously to three Metathesaurus concepts, which have semantic types 'Biologic Function' ("Immunology <1>"), 'Biomedical Occupation or Discipline' ("Immunology <2>"), and 'Laboratory Procedure' ("Immunology <3>"). This experiment thus tests the viability of the rules relevant to these three semantic types.

Our disambiguation rules resolved 86 of the 110 instances correctly (78.2%). The program resolved *immunology* to 'Biologic Function' ("Immunology <1>") 95 times, 75 of which were correct (78.9%). Four instances were resolved to 'Biomedical Occupation or Discipline' ("Immunology <2>"), two of which were correct; and eleven instances were resolved to 'Laboratory Procedure' ("Immunology <3>"), nine of which were correct (81.8%).

CONCLUSION

In conclusion, we would like to suggest that the methodology for ambiguity resolution which we propose can make a significant contribution to increased precision in an information retrieval system. Additional research which we have recently conducted supports this conclusion. As noted earlier this methodology is part of a general algorithm for mapping free text to the UMLS Metathesaurus. In [20] we describe an experiment to test our general mapping algorithm with regard to retrieval effectiveness.

In that study we report that our mapping algorithm chooses the correct Metathesaurus concept around 90% of the time, without word sense disambiguation. Document retrieval conducted on the basis of that mapping resulted in about 60% average precision.

Since incorrect mappings detract from precision and since a significant number of the incorrect mappings are due to ambiguity, resolving ambiguity will contribute to increased precision. Given that the work described in this paper suggests that our method for resolving ambiguous mappings is effective around 80% of the time, we feel that this method shows considerable promise for continued research aimed at increasing precision in information retrieval systems.

References

1. Lindberg DAB, Humphreys BL and McCray AT. "The Unified Medical Language System." *Methods of Information in Medicine* 32:281-291, 1993.

2. Johnson SB, Aguirre A, Peng P and Cimino J. "Interpreting natural language queries using the UMLS." In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 294-298, 1993.

3. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A and Srinivasan S. "UMLS knowledge for biomedical language processing." *Bulletin of the Medical Library Association* 81:184-194.

4. Rindflesch TC and Aronson AR. "Semantic processing in information retrieval." In Safran C (ed.) *Proceedings of the 17th Annual SCAMC*, 611-615, 1993.

5. Moore GW, Hutchins GM, Boitnott JK, Miller RE and Polacsek RA. "Word root translations of 45,564 autopsy reports in MeSH titles." In Stead WW (ed.) *Proceedings of the 11th Annual SCAMC*, 128-132, 1987.

6. Elkin PL, Cimino JJ, Lowe HJ, Aronow DB, Payne TH, Pincetl PS and Barnett GO. "Mapping to MeSH: The art of trapping MeSH equivalence from within narrative text." In Greenes RA (ed) *Proceedings of the 12th Annual SCAMC*, 185-190, 1988.

7. Canfield K, Bray B, Huff S and Warner H. "Database capture of natural language echocardiographic reports: a Unified Medical Language approach." In Kingsland LC, III (ed.) *Proceedings of the 13th Annual SCAMC*, 559-563, 1989.

8. Chute CG, Yang Y, Tuttle MS, Sherertz DD, Olson NE and Erlbaum MS. "A preliminary evaluation of the UMLS Metathesaurus for patient record classification." In Miller RA (ed.) *Proceedings of the 14th Annual SCAMC*,161-165, 1990.

9. Hersh WR, Hickam DD and Leone TJ. "Words, concepts, or both: Optimal indexing units for automated information retrieval." Frisse ME (ed.) *Proceedings of the 16th Annual SCAMC*, 644-648, 1992.

10. Lin R, Lenert L, Middleton B and Shiffman S. "A free-text processing system to capture physical findings: Canonical phrase identification system (CAPIS)." In Clayton PD (ed) *Proceedings of the 15th Annual SCAMC*, 168-172, 1991.

11. Wagner MM. "An automatic indexing method for medical documents." In Clayton PD (ed.) *Proceedings of the 15th Annual SCAMC*, 1011-1017, 1992.

12. Miller RA, Gieszczykiewicz FM, Vries JK and Cooper GF. "CHARTLINE: Providing bibliographic references relevant to patient charts using the UMLS Metathesaurus knowledge sources." In Frisse ME (ed.) *Proceedings of the 16th Annual SCAMC*, 86-90, 1992.

13. Browne AC, McCray AT and Srinivasan S. *The SPECIALIST Lexicon.* National Library of Medicine, Report No. NLM-LHC-93-01 (available from NTIS, Springfield VA: PB93-217248), 1993.

14. Cutting D, Kupiec J, Pedersen J and Sibun P. "A practical part-of-speech tagger." In *Proceedings of the Third Conference on Applied Natural Language Processing*, 1992.

15. Wilks YA. "An artificial intelligence approach to machine translation." In Schank RC and Colby KM (eds) *Computer Models of Thought and Language*, 114-151. San Francisco: W. H. Freeman and Co., 1973.

16. Hirst GJ. *Semantic interpretation against ambiguity.* Brown University Doctoral dissertation, 1984.

17. Stallard D. "The logical analysis of lexical ambiguity." In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 179-185, 1987.

18. McCroy SW. "Using multiple knowledge sources for word sense disambiguation." *Computational Linguistics* 18:1-30, 1992.

19. Schuyler PL, McCray AT and Schoolman HM. "A test collection for experimentation in bibliographic retrieval." Barber B, Cao D, Qin D and Wagner G (eds.) *MEDINFO 89*, Amsterdam: North-Holland, 810-912, 1989.

20. Aronson AR, Rindflesch TC, Browne AC. "Exploiting a large thesaurus for information retrieval." Proceedings of RIAO-94, to appear.